CollegeBoard
connect to college success™

Research Report
No. 2004-3

Beyond Individual Differences: Exploring School Effects on SAT® Scores

Howard T. Everson and Roger E. Millsap

www.collegeboard.com

College Board Research Report No. 2004-3

Beyond Individual Differences: Exploring School Effects on SAT® Scores

Howard T. Everson and Roger E. Millsap

College Entrance Examination Board, New York, 2004

An earlier version of this report was presented as the first author's presidential address to the American Psychological Association's Division of Educational Psychology, August 2000, in Washington, D.C.

Howard T. Everson is vice president for Academic Initiatives and chief research scientist, the College Board.

Roger E. Millsap is professor, Psychology Department, Arizona State University.

Researchers are encouraged to freely express their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

The College Board: Connecting Students to College Success

The College Board is a not-for-profit membership association whose mission is to connect students to college success and opportunity. Founded in 1900, the association is composed of more than 4,500 schools, colleges, universities, and other educational organizations. Each year, the College Board serves over three million students and their parents, 23,000 high schools, and 3,500 colleges through major programs and services in college admissions, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®). The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities, and concerns.

For further information, visit www.collegeboard.com.

Additional copies of this report (item #040481303) may be obtained from College Board Publications, Box 886, New York, NY 10101-0886, 800 323-7155. The price is $15. Please include $4 for postage and handling.

Copyright © 2004 by College Entrance Examination Board. All rights reserved. College Board, Advanced Placement Program, AP, SAT, and the acorn logo are registered trademarks of the College Entrance Examination Board. Connect to college success and SAT Reasoning Test are trademarks owned by the College Entrance Examination Board. PSAT/NMSQT is a registered trademark of the College Entrance Examination Board and National Merit Scholarship Corporation. All other products and services may be trademarks of their respective owners. Visit College Board on the Web: www.collegeboard.com.

Printed in the United States of America.

Contents

Abstract 1

Introduction 1

Overview of Multilevel Latent-Variable Models. 2

Method 3

Student-Level Data 3

Background Data 3

School-Level Data 5

Our Multilevel Modeling Approach 5

Results 5

Stage One: Student-Level Analyses 5

Model Specification 5

Parameter Estimates 8

Stage Two: School-Level Analyses 9

Parameter Estimation 11

The Within-School Model 11

The Between-School Model 13

Discussion 14

Summary 15

Conclusion 16

References 16

Appendix 18

Tables

1. SAT® Verbal and Mathematical Scores by Gender and Race/Ethnicity 3
2. Number of Students Responding to the Survey by Gender and Ethnicity 4
3. High School Achievement Variables 4
4. Extracurricular Activities Variables 4
5. Family Socioeconomic Background Variables 4
6. Fit Statistics for Competing Structural Student-Level Models 7
7. Estimated Slope Intercepts from Model 4 7
8. Intraclass Correlation (ICC) Estimates from Multilevel Model 11
9. Estimates for Ethnic/Gender Paths to SAT Scores of the Within-School Model 11
10. Standardized Path Coefficient Estimates for Latent Predictors of SAT for the Within-School Model 12
11. Standardized Factor-Loading Estimates for the Within-School Model 12
12. Standardized Estimates for Ethnic-by-Gender Paths to Latent Variables for the Within-School Model 13
13. Standardized Path Coefficient Estimates for the Between-School Model 13
14. Standardized Factor-Loading Estimates for the Between-School Model 13

A1. Means of Measured Variables by Race/Ethnicity: Males 18

A2. Means of Measured Variables by Race/Ethnicity: Females 18

Figures

1. Measurement model of the three latent variables at the student level 6
2. Structural model of relationship among measures of family background, high school achievement, and extracurricular activities on SAT verbal and math scores 6
3. Path model of effects of family background, high school achievement, and extracurricular activities on SAT verbal and math scores 8
4. Measurement model of school-level effects . . . 9
5. Within-school model 9
6. Between-school model 10

Abstract

This paper explores the complex, hierarchical relationship among school characteristics, individual differences in academic achievement, extracurricular activities, and socioeconomic background on performance on the SAT Reasoning Test™ verbal and mathematical sections. Using multilevel structural equation models (SEMs) with latent means, data from a national sample of college-bound high school students were analyzed. A nested series of structural equation models were fit simultaneously to eight subgroups (disaggregated by both gender and ethnicity) of high school students. Analyses suggest that multilevel structural equation models provide a reasonably good fit to the data, that family background influences SAT® scores directly and indirectly, that learning opportunities in and outside the school curriculum are related to SAT performance, and that the characteristics of the schools matter when it comes to performance on the SAT. The paper's main contention is that context matters and that researchers ought to move beyond analyses of individual differences when attempting to understand performance on large-scale standardized tests.

Introduction

In his award winning book, *Savage Inequalities: Children in America's Schools*, Jonathan Kozol writes of being startled by the "remarkable degree of racial segregation that persisted…and was common in the public schools" (Kozol, 1991, p. 3). Research on schooling in the United States adds to the bleak picture painted by Kozol, indicating that "African Americans, Hispanics, and American Indians—compared with whites, Asians, and Pacific Islanders—are more likely to attend lower-quality schools with fewer material and teacher resources, and are more likely to have lower test scores, drop out of high school, not graduate from college, and attend lower-ranked programs in higher education" (Dabady, 2003, p. 1048). It should not be too surprising, therefore, to find that minority students, particularly African American and Hispanic youngsters, lag behind on standardized test scores. Indeed, achievement test scores from the National Assessment of Educational Progress (U.S. Department of Education, NCES, 2002), as well as a number of other large-scale assessment programs like the SAT, indicate that African American and Hispanic students score much lower than their white and Asian American classmates in reading, mathematics, and science (College Board, 2003; National Center for Education Statistics, 2002). A host of other academic achievement indicators show similar gaps (Camara and Schmidt, 1999; Jencks and Phillips, 1998; Mickelson, 2003).

Many policymakers believe that reducing or eliminating the achievement gap would go a long way toward reducing racial inequality in the U.S. (Gordon, 1999). For example, when superintendents of large urban school districts were surveyed recently, they listed the issue of the achievement gap between minority and nonminority students as one of their major concerns (Huang, Reiser, Parker, Muniec, and Salvucci, 2003).

Educators, however, are perplexed when it comes to finding affordable ways of raising African American and Latino students' achievement. For decades now, particularly since the publication of the report by James Coleman in 1966, *Equality of Educational Opportunity*, many policymakers and educators have opined that schools—and in particular our public schools—could do little to address the inequities of poverty and the academic underachievement of minority students. It was a matter of underdeveloped academic ability, a view rooted in a strong belief in individual differences. Schooling had little or no effect on standardized test scores.

As a consequence, relatively little effort was directed at understanding subgroup differences—either gender or race/ethnicity—in standardized test scores throughout the latter half of the last century, and even less attention was given to the complex interaction between gender and race (Anderson and Bruschi, 1993; Jones, 1987). The projected demographic shifts in the United States in the early decades of the twenty-first century will press educational psychologists and other researchers to better understand the complexities of academic achievement and the effects of schooling on standardized test scores. In response, we are seeing an increase in research aimed at identifying effective schools (see, for example, Cohen, Raudenbush, and Ball, 2002; Good and Brophy, 1986; Hanushek, 1997; Lee, Bryk, and Smith, 1993; and Lee, 2000). These efforts are beginning to shift the framework for educational research from asking how schools affect learning to understanding how to maximize resources to promote achievement. What resources matter and what are the benchmarks of sustainable academic achievement? Educators, parents, and policymakers have long assumed that schools and their attendant resources matter when it comes to student achievement. Yet the past two decades of research tell us that the relationship between schooling and student achievement is neither direct, nor easily understood. Schools across the United States are organized in different ways and they use resources differently. The effects of these organizational and structural differences on student achievement are made more complex when we recognize, as we must, the variety of individual differences in background and ability that accompany students in our schools

Relying on advances in statistical methods, in particular multilevel covariance analysis or structural equation modeling, we set out to explore the complex relationships and interactions among individual differences in socioeconomic background, gender and ethnicity, and academic ability as they influence performance on a high-stakes standardized test, the SAT. By introducing multilevel models, we attempted to isolate structural differences in secondary schools—i.e., school effects—on SAT scores. We used a matched comparison group design to develop our evidence. Our research goals were (1) to better understand the correlates of the SAT score gaps across subgroups of examinees based on gender and race/ethnicity; (2) to develop explanatory models (single and multilevel) that fit reasonably well the SAT background and performance data across all subgroups; and (3) to identify the differential effects of family background, academic achievement, and school-level variables on SAT performance.

The remainder of this report is divided into four sections. To begin, a brief overview is provided of multilevel latent-variable modeling methods, arguing for their particular advantages when complex, multivariate, multilevel questions are addressed. The second section describes the data analyzed in this study—highlighting when and how the data were gathered, and how the data were structured to model latent variables. The third section of the report provides detail on the results of the model fitting at both the single and multilevel stages. The report concludes with a discussion of what multilevel models suggest about the relative contributions of individual differences in students' backgrounds, abilities, academic experiences, and secondary school characteristics to performance on the SAT verbal and mathematical reasoning tests.

Overview of Multilevel Latent-Variable Models

Over the past decade or so, multilevel modeling techniques have been developed to investigate school effects, which, by their nature, are hierarchical (or multilevel) because students are nested within schools. "School effects," then, refers to influences of school qualities and characteristics on student achievement, e.g., test scores, or other educational outcomes. At the time when Coleman (1966) was studying school effects and until the mid-1980s, it was common to investigate school effects using statistical regression methods or analysis of variance. There are obvious conceptual and methodological difficulties in using these methods, not the least of which was the troublesome confounding of variability at the individual student level with the variation within and between aggregated levels (e.g., schools) of the data. (For a more technical discussion of these difficulties, see Bryk and Raudenbush, 1992; Goldstein, 1987; Lee, 2000; Millsap, 2002; and Snijders and Bosker, 1999.) To resolve these difficulties, multilevel models were developed that allowed for the variance attributable to the school level to be partitioned from the variance associated with individual differences in students, permitting the estimation of more accurate, less biased, standard errors and more reliable and useful information about between- and within-school effects.

More recently, multilevel structural equation modeling (SEM) has become a popular methodology for studying complex phenomena in the social and behavioral sciences, and is a vigorous line of methodological research (Raudenbush and Sampson, 1999). Generally speaking, SEMs are simultaneous equation models that permit researchers to build and test models containing both endogenous and exogenous latent variables, which, as in factor analysis, can be represented by multiple indicators, i.e., multiple observed variables. Within this framework, SEMs combine both measurement and structural models. The measurement model sets forth the relations (i.e., factor loadings) between the latent constructs and their observed indicators, along with the unique or error variance associated with each observed variable. The structural model, in contrast, describes the directional structural relationships among the latent variables. Longford (1993) provided maximum likelihood estimation methods for two-level regression models with latent variables, each with multiple indicators. Others, e.g., Muthen (1984), Muthen and Muthen (1998), and McDonald, (1993), extended these estimation methods to address problems of missing data at either level, and made available specialized software for conducting multilevel covariance structure analysis.

This family of statistical techniques blends path analysis and factor analysis in a framework for assessing causal models (Hoyle, 1995; Loehlin, 1992). As such, SEM is a very general, largely linear statistical modeling technique, and a powerful tool for representing a network of hypothesized linear relations among a complex set of variables aggregated or nested at multilevels. This approach is particularly well suited for our study because of the large number of observed variables in our model, and our interest in linking performance on the SAT with school characteristics.

In SEM each measured variable is expressed as a linear function of one or more common factors, and a single unique factor. The common factors include influences on the measured variables that are shared among two or more such variables. For example, in the set of questions designed to measure achievement in high school across a variety of academic subjects, we hypothesized that students' responses share a single common factor: academic achievement. The unique factors include influences that are specific to each measured variable, such as random measurement error, or systematic components such as method influences.

Both the common and unique factors are denoted as latent variables in SEM, but interest focuses mainly on the common factors, and so references to latent variables often are meant to apply to the common factors. Relations among the common factors are, in turn, represented by a set of path equations that capture the hypothesized causal paths among these latent variables. The common factors are intended to represent the constructs of interest, in this case academic achievement, extracurricular activities, socioeconomic background, and SAT scores.

The equations formally resemble regression equations, with each equation expressing the modeled value of a criterion or *endogenous* variable as a linear function of one or more predictor variables, plus a residual or disturbance term. Unlike ordinary regression equations, however, path model equations are conceived as explicitly causal, with the regression or path coefficients representing the direct causal influence of the predictor variable on the endogenous variable. In path analysis, predictor variables may themselves be *endogenous* in relation to other predictors, or may serve purely as predictors. Variables of the latter type are denoted as *exogenous* variables in path analysis, while all other variables are regarded as endogenous. The path model represents a causal theory for the endogenous variables. Causal influences on the exogenous variables are not directly modeled. The challenge facing us was to specify a model that expressed how extracurricular academic activities, family background, and academic achievement all influence performance on the SAT—and determine if the hypothesized model represents performance across all subgroups of students—while simultaneously examining the independent effects of the schools' qualities and characteristics on test performance.

Method

As we noted earlier, this study was animated by a concern about the persistent achievement gap between minority and nonminority students on large-scale standardized tests like the SAT, and by related worries that perhaps standardized tests like the SAT are biased measures of cognitive ability. Always mindful of the policy issues related to school improvement, accountability, and the achievement gap, we wanted to disentangle the role of individual differences in academic achievement from other possible contributors—including, for example, socioeconomic status, educational opportunities, and school effects—on test performance. Using data collected from the College Board's *Student Descriptive Questionnaire*, which is administered annually to students taking the SAT, we developed and tested a series of multilevel latent-variable models and fit them to the SAT data. We attempted to fit our complex model to eight groups of students—the male-female subsets of African Americans, Hispanics, Asian Americans, and whites. Specifically, we asked whether the observed score differences on the SAT remained after controlling for family background, course-taking opportunities, and academic achievement. Moreover, we asked if the introduction of high school characteristics, e.g., size of the school, the percent minority, its location (urban, rural, suburban), and the percent of students eligible for free or reduced lunch (an indicator of a school's socioeconomic status) would contribute to the further reduction of group differences on SAT scores.

Student-Level Data

Our data come from a subset of college-bound seniors who took the SAT during their junior or senior year of high school, and who graduated from high school in 1995. This cohort of 1.14 million students had mean SAT verbal and mathematical scores of 504 and 506, respectively, on the recentered SAT scale (Dorans, 2002). They represent about 41 percent of all the high school seniors in the United States in 1995. Girls make up about 54 percent of this group, and the cohort is largely white (69 percent), with 11 percent African American, 8 percent Asian American, 4 percent Mexican American, 4 percent other Latinos, 1 percent Native American, and 3 percent who marked "other" when noting their race or ethnicity. Because our analyses focus on subgroup differences in SAT scores, Table 1, below, displays the mean SAT verbal and mathematical scores disaggregated by race/ethnicity and gender for this cohort of college-bound students.

The magnitude of group differences in SAT scores is clear. Males outperform females in mathematics, and white and Asian American students, in general, score higher on both the verbal and mathematical SAT tests than African American and Hispanic students.

Background Data

When students register with the College Board to sit for the SAT, they complete a lengthy questionnaire, answering 43 questions about their high school courses,

Table 1

SAT Verbal and Mathematical Scores by Gender and Race/Ethnicity

	Whites	*Asian Americans*	*African Americans*	*Hispanics*
Males				
SAT–M	551	577	444	495
SAT–V	537	533	443	484
Females				
SAT–M	515	543	427	462
SAT–V	534	531	448	478

participation in a sweep of extracurricular activities, academic achievement levels (i.e., grades), parental education, family income, and their race or ethnicity (see www.collegeboard.com for a copy of the *Student Descriptive Questionnaire*). Responses to these questions formed much of the data for this study. For this study, we excluded students who reported that they were not U.S. citizens, and for whom English was not their first language. In addition, we also excluded students who had not attended a public high school in the United States at the time of testing. We also excluded students who were missing responses on the variables of interest in this study. Table 2 shows the number of students enrolled in public high schools, disaggregated by race/ethnicity and gender, responding to all the relevant questions in the College Board survey, thus comprising our sample. This subset of more than 484,000 students provided the data used in the subsequent analyses.

We chose carefully and empirically the measured variables to serve as indicators of each hypothesized latent construct, recognizing, too, that these self-report measures are imperfect and often unreliable. At this first level, the student level, we were interested in three latent variables—the family socioeconomic background, academic achievement in high school, and academically related extracurricular activities—and the SAT verbal and mathematical scores.

The College Board questionnaire provided the data used to construct the latent variables. For example, students were asked to indicate the total number of years they took high school courses in specific subject areas, and to report their grade point average (GPA) on a scale of A to F for each academic subject. These data elements were used to model academic achievement, and they are presented in Table 3, at right.

Similarly, students indicated their participation in a range of extracurricular activities. Table 4 provides the complete list of variables we used to model participation in academic and nonacademic extracurricular activities while in high school.

Students in our sample also reported their best estimates of annual family income in increments of $10,000, with reporting categories ranging from a low of $10,000 to a maximum of $100,000 or more per year. In addition, they reported the highest level of education attained by both parents. These three variables were used to model students' socioeconomic backgrounds. See Table 5, at right.

And finally, in addition to these self-report measures, each student's SAT verbal and mathematical scores (reported on a scale from 200 to 800) were used as the outcome measures in our analyses. (The means for all measured variables at the student level, broken down by gender and ethnicity, are presented in the Appendix).

Table 2

Number of Students Responding to the Survey by Gender and Ethnicity

	Males	*Females*	*Total*
Whites	170,270	212,412	382,682
African Americans	18,411	27,644	46,055
Asian Americans	12,333	13,732	26,065
Hispanics	13,026	16,666	29,692
Total	214,040	270,454	484,494

Table 3

High School Achievement Variables

HSAVG	High School Grade Point Average
CRANK	High School Class Rank
ARTGR	GPA in Art and Music Courses
SOCGR	GPA in Social Science and History
ENGR	GPA in English Courses
LANGR	GPA in Foreign Language Courses
MATHGR	GPA in Mathematics Courses
SCIGR	GPA in Natural Science Courses

Table 4

Extracurricular Activities Variables

ACTCNT	Number of Extracurricular Activities (pursued for at least 3 years)
APCNT	Number of AP® Exams Intended
HNRCNT	Number of Honors Classes Taken
ENGCNT	Number of Literature Experiences
COMPCNT	Number of Computer Experiences
ARTCNT	Number of Art, Music, and Theater Experiences

Table 5

Family Socioeconomic Background Variables

FATHED	Father's Education Level
MOTHED	Mother's Education Level
FAMINC	Combined Parental Income

School-Level Data

Four school-level variables were selected from the 1994 NCES database for public secondary schools in the United States (National Center for Education Statistics, 1994). These variables were limited in number and were based on preliminary correlations with SAT verbal and mathematical scores. The four selected variables were (1) the number of students in the high school eligible for free or reduced lunch (FLE), a measure of the socioeconomic status of the students at that school; (2) the total number of students in the high school (MEMBER), an index of the size of the school; (3) the proportion of minority students in the school (PNWHIT); and (4) the location of the school (LOCALE) on a seven-point scale ranging from rural to suburban to urban. These school-level characteristics were merged with the student-level data using common codes on both the College Board and NCES databases, resulting in a database of 8,258 high schools and 288,066 students nested within those schools and available for multilevel analysis.

Our Multilevel Modeling Approach

As we said earlier, our analytic approach relied on multilevel structural equation modeling (SEM). This approach is particularly well suited for this study because of the large number, 19 in all, of observed variables in our model, and our interest in linking student-level data, as well as school-level data, with performance on the SAT. The structural equation modeling of the combined student and school data proceeded in two broad stages. In the first stage, a nested series of structural equation models were fit to the student-level data, omitting the school-level data. For these single-level analyses, multiple student groups were created based on ethnic group membership and gender, yielding eight groups (Asian American males and females, Hispanic males and females, African American males and females, and white males and females). All models in this stage were fit simultaneously in these eight groups, permitting tests of ethnic and gender invariance of the path coefficients in the structural and measurement models. Mean structures were included in all models. The use of means permitted tests for group differences in mean SAT performance after adjustment for modeled influences on the SAT.

The second stage examined a multilevel structural equation model that included influences of the school-level variables on SAT performance at the school level, using nearly the same student-level structural model as in the first stage. The new feature was that ethnic and gender differences on the latent variables were specified as direct effects using contrast-coded group definitions for the ethnic and gender subgroups. Hence, these were not simultaneous analyses using multiple groups defined by ethnicity and gender. Instead, variables were created to represent ethnic group membership, gender, and the interaction of ethnicity and gender. These variables were then included in the model as exogenous measured variables. This model structure permitted an evaluation of ethnic and gender effects on SAT performance within a school after adjustment for the modeled influences on the SAT.

Results

Here we describe the two modeling stages that characterize our study. We also include descriptions of the models tested and present our preliminary conclusions. We begin with stage one, the student-level model.

Stage One: Student-Level Analyses

At this first level we examined the relationships among and influences of socioeconomic background, academic achievement, and extracurricular activity levels on high school students' verbal and mathematical SAT scores. We looked at these relationships across eight subgroups of students, based on ethnic group membership and gender. Our explanatory models were developed and tested against the SAT verbal and mathematical scores of students in all eight subgroups. We proceeded in three broad stages: (1) specifying a model that relates the variables to one another; (2) estimating the parameters of the model; and (3) estimating how well the model fits the empirical data, that is, how well the theoretical model replicates the empirical correlations between and among the variables included in our database.

Specifying the model required us to translate the theory we wished to test, in this case the relationship between latent variables and SAT scores, into a particular structural model that could be derived and tested given the empirical data on hand. Thus, the resulting models, we hoped, would not be refuted by our data. At the parameter estimation stage we used the College Board data to derive estimates of the model parameters—i.e., the coefficients calibrating the relationships among the variables—deemed optimal by one or more statistical estimation methods. And finally, to evaluate the fit of our model, we used the derived parameter estimates to examine how well the hypothesized model reproduced the covariation found in the empirical data.

Model Specification

This step began with a theory about the relationships among the variables under study. For convenience, a distinction was made between the *measurement* and *structural* portions of the model. The measurement model describes the relationship between the measured variables and the latent variables—the underlying factors that account for the hypothesized relationships among the observed or measured variables. Thus, the 19 measured variables in our study were tested empirically and were represented, ultimately, by three latent variables—socioeconomic background, high school achievement, and extracurricular activities—and two observed variables (the SAT verbal and mathematical test scores). The measurement model for the three latent variables is depicted in Figure 1.

The boxes in Figure 1 represent measured variables from the College Board questionnaire completed by all the students taking the SAT, while the circles represent the common factors or latent variables hypothesized to underlie the 17 measured variables (excluding the two SAT scores). The directed arrows linking the latent and measured variables indicate which observed variables were hypothesized as measures of each latent variable. In our model, each measured variable was linked to a single latent variable.

The *structural* portion of the model specified the directional relations among the latent variables, or among the measured variables (the SAT scores) when no latent variables are included. The choice of which latent variables are linked directly by paths, and which are related indirectly, is based on theory and findings from our earlier research (see Everson, Millsap, and Diones, 1995). Thus, our first consideration when defining the structural model was the choice of which latent variables are exogenous, and considered causally prior to the other latent variables. Our model asserts that performance on the SAT is dependent upon high school achievement and participation in extracurricular activities (Marsh, 1992), both of which, in turn, are dependent on socioeconomic background. We also hypothesized that socioeconomic background influences directly SAT scores (Everson, et al., 1995; Everson and Michna, 2004). This hypothesized structural model is depicted in Figure 2.

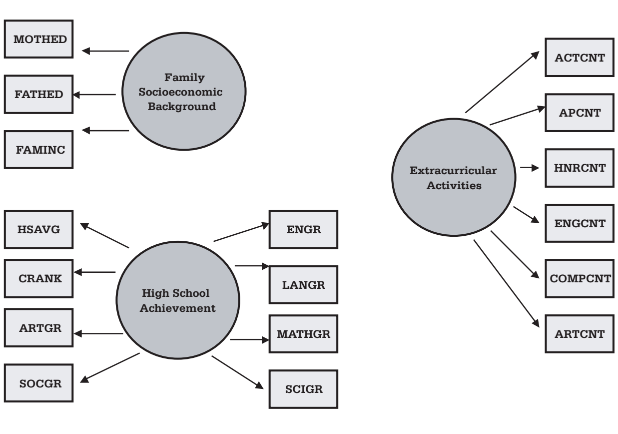

Figure 1. Measurement model of the three latent variables at the student level.

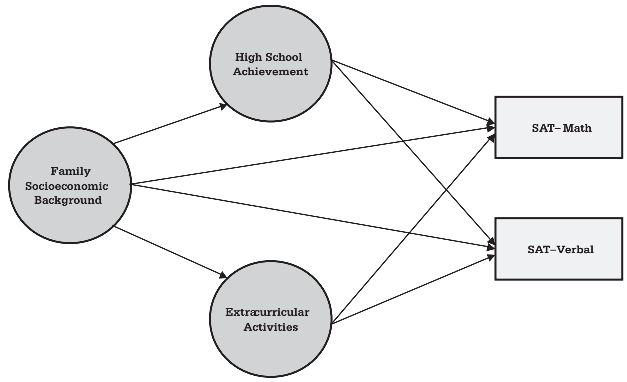

Figure 2. Structural model of relationship among measures of family background, high school achievement, and extracurricular activities on SAT verbal and math scores.

Again, the SEM approach centers on two steps: validating the measurement model and fitting the structural model. We divided our sample of N = 484,494 into eight groups based on the factorial combination of gender with the four ethnic classifications: Asian American, African American, Hispanic, and white. In this initial stage, a nested series of structural equation models were fit to the data, permitting tests of ethnic and gender invariance of the structural and measurement models. Mean structures were used to allow for tests for group differences in mean SAT verbal and mathematical performance after adjusting for the influences of the other factors. The use of means permitted tests for group differences in average SAT performance after adjustment for modeled influences—socioeconomic background, achievement, and extracurricular activities—on the SAT scores. We proceeded, then, by fitting a series of measurement and structural models, starting with the most general—least constrained—model and moving on to more specific models that assumed invariant or constrained relationships across all subgroups of students.

Our initial model, Model 1, fit a five-factor model to the 19 measured variables within each of the eight student groups. This is the most general model, and in some ways the least interesting, though it does suggest that the underlying factors do account for the relationships among the observed variables. In Table 6 we present the fit information for Model 1, and for all subsequent models we fit to these data. Three fit statistics are given for each model at the student level: the chi-square fit statistic and degrees of freedom, the root mean square error of approximation (RMSEA) (Steiger and Lind, 1980), and the comparative fit index (CFI) (Bentler, 1990; Bentler and Bonett, 1980; McDonald and Marsh, 1990).

The “null” model in Table 6 is the independence model needed for the calculation of the various fit indices, but is of little intrinsic interest. Model 1, the measurement model, provides a good fit, as indicated by the RMSEA

Table 6

Fit Statistics for Competing Structural Student-Level Models

Model	*Chi-Square*	*df*	*RMSEA*	*CFI*
Null	3,862,416	1,368		
(1) Measurement model, unconstrained loadings	87,831	704	.045	.98
(2) Measurement model, congeneric	286,713	1,152	.064	.93
(3) Measurement model, invariant loadings	307,133	1,250	.062	.92
(4) Structural model, invariant loadings, and paths	309,145	1,306	.062	.92
(5) Structural model, invariant loadings, paths, and intercepts	374,822	1,334	.068	.90
(6) Structural model, invariant loadings, paths, partial invar. on intercepts	346,637	1,326	.065	.91

and CFI indices. Again, these fit indices support the claim that five factors are sufficient to represent the 19 measured variables in all groups at the student level.

Model 2 constrains Model 1 by requiring that each measured variable load on one and only one underlying factor—constraining the crossloading of an observed variable to no more than one underlying factor. In Model 2, however, two of the five factors are presumed to have nonzero loadings only for the SAT variables, with one factor representing SAT–V and the other SAT–M.

Model 2, in short, specifies that the observed SAT verbal and mathematical scores represent the latent factors of verbal reasoning and mathematical reasoning. The overall fit is acceptable, as indicated by the RMSEA and CFI indices.

Further analyses suggest that the slight loss of fit in Model 2 relative to Model 1 results primarily from the sharp distinction between the extracurricular activities and high school achievement factors. We suspect, for example, that some of the observed variables in these factors may have nonzero loadings on *both* factors—high school achievement and extracurricular activities—rather than only on one of the two. The variable HNRCNT, for example, which counts the number of honors courses taken, was constrained statistically to load only on the extracurricular activities factor in Model 2, but we expected, nevertheless, that it had nonzero loadings on both the high school achievement and extracurricular activities factors. The results, obviously, suggest that while the academic achievement–extracurricular activities distinction may not be as sharp as we had believed initially, the hypothesized five-factor structure is, nevertheless, a good approximation of the relationships in the data. We will return to this issue later.

Model 3 further constrains Model 2 by forcing the factor loadings to be invariant across all eight subgroups of students. Apart from these invariance constraints, all other parameter matrices and estimates were expected to have the same structure as in Model 2. The constraints introduced in Model 3 did not degrade the fit relative to Model 2, suggesting that the factor loadings (or functional weights) can be presumed to be invariant across groups without substantial loss of fit of the model. Clearly, the fit indices of these first three models provide confidence that a five-factor measurement model with invariant factor loadings fits the data reasonably well.

Next, we fit a series of models that added restrictions on the relationships among the five underlying factors, creating a combined measurement and structural model. Again, see Figures 1 and 2 as representations of these hypothetical relationships.

Model 4 examined the invariance restrictions on the coefficients (the strength of relationship) of the paths among and between the underlying factors. The comparison between Models 3 and 4 is a test of these invariance restrictions, i.e., that the purported causal relationships among the latent factors are more or less the same across all the subgroups. To underscore this view, in Table 7 we present the estimated structural intercepts for SAT–M and SAT–V in each of the eight groups that result from Model 4.

The key comparison is between group differences in the slope intercept estimates in Table 7, and the observed group differences in the corresponding SAT means in Table 1. To illustrate, Table 1 reveals a difference of more than 100 points on SAT–M between white and African American males. Yet, conditional analyses based on the student-level latent variable models, in particular Model 4, suggest that the differences on SAT–M are quite a bit lower, perhaps only 50 points on the SAT scale for these two groups of examinees. This dramatic reduction in SAT–M score differences represents the statistical adjustment for socioeconomic background, high school achievement, and extracurricular activities. That is, *ceteris*

Table 7

Estimated Slope Intercepts from Model 4

	Whites	*Asian Americans*	*African Americans*	*Hispanics*
Males				
SAT–M	205	231	154	200
SAT–V	240	240	198	241
Females				
SAT–M	152	177	120	155
SAT–V	218	214	187	222

parabus, the difference in SAT–M performance drops to 50 points in this sample. The remaining score differences are unexplained after adjustment for these three explanatory latent factors.

To take another example, the SAT–M and SAT–V mean differences between Hispanic and white males are essentially eliminated by the adjustment for the explanatory factors. In contrast, the gender differences on SAT–M in Table 1 within each of the white, Hispanic, and Asian groups are smaller than the SAT–M gender difference within these groups after we adjusted for socioeconomic background, high school achievement, and extracurricular activities variables. Here the adjustment for the explanatory factors served to widen the gender difference, rather than eliminate it. The basis for this finding appears to lie in the complex pattern of gender differences on the socioeconomic background, high school achievement, and extracurricular activities variables in Table 1. Females score higher on the high school achievement variables, and they score higher on most of the extracurricular activities variables. These results suggest that after adjusting for the females' higher scores on these latent variables, we expect to see even larger gender differences on the SAT–M than we observe in the unadjusted, larger population of males and females. The higher academic performance of the females in the unadjusted population serves to reduce the average gender difference on SAT–M. Once this higher academic performance is attenuated via statistical adjustment, the SAT–M score difference in favor of males—across all ethnic groups—is increased.

Model 5 is identical to Model 4 with the exception that now the structural intercepts (the latent means of the SAT–V and SAT–M scores) are constrained to be invariant across all groups. This restriction suggests that after adjusting for socioeconomic background, high school achievement, and extracurricular activities, there are no group differences in expected or mean SAT scores. If Model 5 fits well, we may be able to explain the observed group differences in SAT performance in terms of group differences on the three underlying factors in our theory-based structural model. Obviously, Model 5 is particularly important for this reason, and its fit must be carefully examined.

Table 6 shows that while there is some global loss of fit associated with the invariance restrictions on the structural intercepts noted above, the overall fit remains reasonably good. A more detailed look at the fit of Model 5 suggested that it does not fit perfectly, but the important question is whether the fit is below an acceptable threshold in any of the groups. Further inspection of the SAT–V and SAT–M means within each of the eight groups revealed that the African American SAT means are substantially lower than would be predicted by Model 5. The discrepancy is around a half of a standard deviation (50 points) for both males and females on both the SAT–M and SAT–V score scales. This discrepancy suggests that while the invariance restrictions imposed in Model 5 may not degrade the fit of the model globally, the restrictions are too stringent for the African American group. The key conclusion is that after adjustment for the socioeconomic background, academic achievement, and extracurricular activities latent variables, the African American students—both males and females—continue to score lower on SAT–V and SAT–M than would be expected by Model 5.

The final model, Model 6, relaxed the invariance restrictions on the structural intercepts for African American males and females in Model 5. The remaining six ethnic-by-gender groups are restricted to invariance on the structural intercepts, as in Model 5. All other parameter restrictions in Model 5 are retained in Model 6. As shown in Table 6, the global fit indices improve slightly with the loosening of the restrictions on the structural intercepts for the African American groups. Given this improvement, Model 6 was preferred over Model 5.

Parameter Estimates

Given the model fitting described in the preceding section, the standardized path coefficient estimates presented here were derived from Model 6. Recall our aim was to derive a global standardization method for estimating a common metric across the eight groups (with the caveats on the structural intercepts for the African American students noted earlier), permitting the creation of a single set of standardized estimates. The comparative strengths of these path coefficient estimates are presented in Figure 3.

Overall, and somewhat surprisingly, we see that the direct influence of students' extracurricular activities on their SAT–V and SAT–M scores is larger than the influence of their academic achievement levels. We found, for example, that a unit change, a standard deviation difference, in the extracurricular activities latent variable

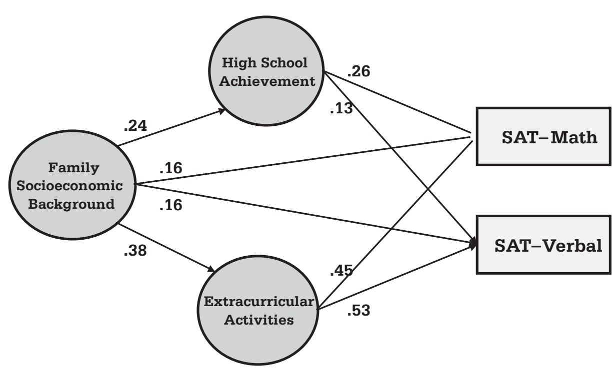

Figure 3. Path model of effects of family background, high school achievement, and extracurricular activities on SAT verbal and math scores.

produces a 45-point increase in SAT mathematical scores, and a 53-point change in SAT verbal scores (i.e., roughly a half standard deviation increase). In contrast, a unit change in socioeconomic status (roughly equivalent to a $20,000 increment in family income) only results in a 16-point increase in SAT verbal and mathematical scores. Although the direct influence of students' socioeconomic backgrounds is relatively small, there is an appreciable indirect relationship to SAT performance, with family background influencing achievement and exposure to extracurricular activities which, in turn, have direct bearing on test scores. The squared multiple correlations for SAT–V (R^2 = .49) and SAT–M (R^2 = .57), an index of the explanatory power of the structural model, suggest that the structural model provides a reasonably good fit to the SAT scores. Again, although the squared multiple correlations are somewhat lower for African Americans (in the range of .35 to .45), the structural model represented by Figure 3, in general, accounts for about half of the variance in the SAT scores at the student level.

Stage Two: School-Level Analyses

As we noted earlier, in an attempt at estimating school effects, four school-level variables were selected from the NCES database for public secondary schools, only. The four variables were then merged with the student-level data, resulting in a database of 288,066 students nested within 8,258 high schools. The model of school-level effects is represented in Figure 4, below.

Unlike the student-level analyses discussed earlier, the multilevel analysis of school effects combined the eight ethnic-by-gender groups into a single group. Ethnic and gender group membership, then, was represented by a set of effect codes or contrasts among the eight groups. Three contrasts were used to represent the four ethnic groups, and a single contrast was used to represent gender. The gender contrast assigned males a value of 1 and females a value of -1. This contrast was labeled SEX. The first ethnic contrast (ETH1) involved whites and African Americans: white = 1, African American = -1, Asian Americans and Hispanics = 0. The second contrast (ETH2) compared whites and Asian Americans: whites = 1, Asian Americans = -1, African Americans and Hispanics = 0. The third ethnic contrast (ETH3) compared whites and Hispanics: whites = 1, Hispanics = -1, African Americans and Asian Americans = 0. These ethnic and gender contrasts represented overall group differences, either among ethnic groups or between males and females. The combined effects of ethnicity and gender (i.e., interactive influence of ethnicity and gender) were represented by creating three additional contrasts as the cross-products of the gender contrast with each of the three ethnic contrasts described earlier. For example, the first cross-product multiplied the gender contrast by the first ethnic contrast (i.e., whites = 1, African Americans = -1). This created a new contrast (SETH1) that compared the difference between white males and females to the difference between African American males and females. The remaining two cross-products (SETH2 and SETH3) have similar interpretations. In all, seven contrast variables were created to represent the effects of gender and ethnicity, and their interactions.

The school effects model used for the combined student and school data consisted of two models, one fit to the individual students within a given school, and the other fit to the aggregate data at the school level. In what follows, the former model is denoted as the "within-school" model, and the latter denoted as the "between-school" model. In this scheme, the school-level variables (i.e., FLE, MEMBER, PNWHIT, LOCALE) were included in the between-school model as exogenous variables, but were omitted from the within-school model. Both models included the ethnic/gender contrast variables, but only the within-school model included constraints on the relationships between these variables and the remaining variables measured on the students. Figure 5 displays the structural portion of the within-school model.

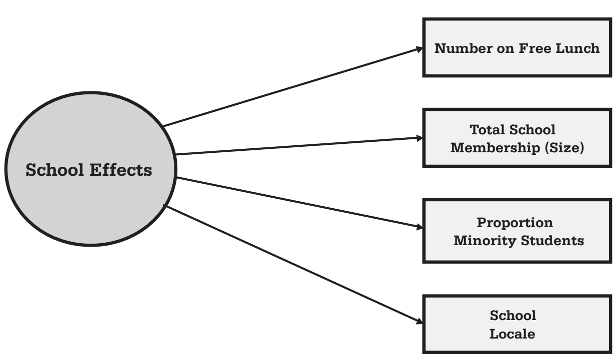

Figure 4. Measurement model of school-level effects.

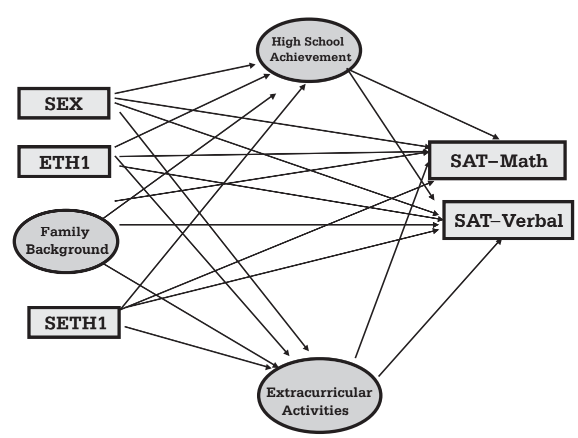

Figure 5. Within-school model.

This structural model in Figure 5 is identical to the model used in stage one at the student level, except with regard to the role of ethnicity and gender. Ethnicity and gender are represented by the seven contrast variables, which are exogenous variables in the model. In Figure 5, the seven contrast variables are reduced to three to simplify the presentation. One variable represents the three ethnic contrasts, one variable represents the gender contrast, and one variable represents the three interaction contrasts. Direct paths are included from the ethnic/gender variables to school achievement, school activities, and the two SAT performance variables. These direct paths represent ethnic/gender group differences in the endogenous variables in the model.

The direct paths from the ethnic/gender contrasts to SAT–M and SAT–V are of central importance. If the path coefficients are insignificant, for example, we would conclude that there are no meaningful ethnic or gender differences in SAT performance after adjusting for family background, high school achievement, extracurricular activities, and school membership itself. The adjustment for school membership follows from the within-school aspect of the model. On the other hand, if at least some of the direct paths from the ethnic/gender contrasts to SAT–M and SAT–V are meaningful, this would be evidence of group differences in SAT–M and/or SAT–V that are not explained on the basis of family background, high school achievement, extracurricular activities, or the school attended. In this sense, the direct paths play a role in the model that is similar to the role of between-group differences in the structural intercepts in the earlier, student-level, stage-one models. It is also worth noting that the measurement portion of the within-school model of the three latent variables of family background, high school achievement, and extracurricular activities is the same as the measurement model used in stage one, as shown earlier in Figure 1. The same observed indicators were used in both stages of our analyses.

In keeping with our general modeling approach, the between-school structural model is displayed in Figure 6. At this level, our between-school model replicates a portion of the within-school structural model, but includes the school-level variables as exogenous variables with direct paths to school achievement, school activities, and SAT performance. The ethnic/gender contrast variables were also included in the between-school model, but were given no directional relations to any of the other variables in the between-school model. In effect, the model places no constraints on the relations between the ethnic-by-gender contrasts and the other variables in the model. Hence the contrast variables included in Figure 5 are omitted from Figure 6.

The between-school model, then, makes no attempt to study ethnic/gender differences at the school level, apart from the inclusion of the percentage of minority students (i.e., the PNWHIT variable) as a school-level exogenous variable. The between-school model, however, does provide information about the roles of the three within-school latent variables, i.e., family background, high school achievement, and extracurricular activities, in relation to SAT performance. In the between-school model, all of these measures are, in effect, considered at the aggregate level by school. For example, SAT–M now represents the average SAT–M score at a given school. Similarly, family background represents an average family background for the students enrolled in a particular school. (Note that for the latent variables of family background, high school achievement, and extracurricular activities, the aggregate scores are based on those students from the school who appear in the database, rather than on the full student population at the school.) Obviously, in any given school, only a portion of the student population will take the SAT, and only a portion of those students appear in the database, depending on when the test was taken or on missing data. Moreover, the measurement portion of the between-school model in relation to the three latent variables was identical to the measurement used in the within-school model. The combined multilevel model pictured in Figure 6 was fit to the data using *Mplus* software, version 1.04 (Muthen and Muthen, 1998). Again, this model fit the data reasonably well (X^2 = 161710.055, df = 540, and RMSEA = .032).

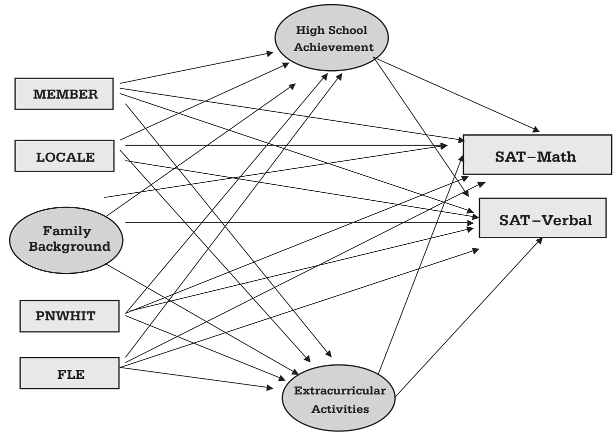

Figure 6. Between-school model.

To aid understanding of between-school effects, it is often helpful to examine the intraclass correlations (the degree of association between the school-level and student-level measures). Table 8 presents the intraclass correlation (ICC) estimates for the full set of 19 measured student-level variables. These correlations indicate the relative size of the between-school variation in each variable to the total variation in that variable. Thus, a high correlation suggests that variation between schools in that variable is high in comparison to the variation of that same measure within schools.

Table 8

Intraclass Correlation (ICC) Estimates from Multilevel Model

Variable	*ICC*
HSAVG	.110
CRANK	.059
ARTGR	.054
SOCGR	.100
ENGR	.106
LANGR	.076
MATHGR	.087
SCIGR	.094
ACTCNT	.068
APCNT	.072
HNRCNT	.146
ENGCNT	.069
COMPCNT	.090
ARTCNT	.043
FATHED	.171
MOTHED	.111
FAMINC	.191
SATM	.164
SATV	.145

Not surprisingly, the highest ICC estimates are for two of the three family background measures (FATHED and FAMINC). This result is likely explained by variations in socioeconomic conditions across school districts, which, in turn, are associated with the socioeconomic status of families of students attending a given school. The two SAT performance variables also have ICC estimates that are fairly large in comparison to those of most of the other measured variables.

Parameter Estimation

The Within-School Model. We begin the discussion of the parameter estimates with the within-school model, and the relationship between ethnicity, gender, and the SAT performance variables. Table 9 gives the raw path coefficient estimates for the direct paths from the ethnicity and gender contrasts to SAT–M and SAT–V.

The raw or unstandardized metric was chosen for these categorical predictors because it is translated directly into group differences on the SAT scale, adjusting for other predictors in the model. Consider, for example, the estimate for the path from SEX to SAT–M of 21.8 in Table 9. This estimate indicates that the difference between the male and female SAT–M means, after adjustment for the other predictors of SAT–M, is (2)(21.8) = 43.6 points on the SAT–M scale. This is a few points less than one-half of a standard deviation on the SAT–M scale, with males having the higher SAT–M scores. The result suggests that even after adjusting for the other predictors of SAT–M (ethnicity, family background, high school achievement, and extracurricular school activities) and for the school attended, there remains about one-half of a standard deviation difference between males and females in performance on the SAT–M.

Table 9

Estimates for Ethnic/Gender Paths to SAT Scores of the Within-School Model

	SAT–M	*SAT–V*
SEX	21.8	6.6
ETH1	19.7	12.5
ETH2	-9.2	10.6
ETH3	2.7	-4.8
SETH1	2.6	0.3*
SETH2	-0.3*	0.9*
SETH3	0.1*	-1.0*

* Not statistically significant at alpha = .05.

Turning to the ethnic results, the coefficient estimate for ETH1 is 19.7. The coefficient estimates for the three ethnic predictors (ETH1, ETH2, ETH3) can be interpreted as the difference between the average of the ethnic group SAT–M means, and the particular minority group involved. For example, ETH1 involves the African American group. The estimate 19.7 indicates that the African American group SAT–M mean is 19.7 points below the average of the ethnic group means, after adjusting for the other predictors of SAT–M. The estimate of -9.2 for ETH2 indicates that the Asian group is 9.2 points above the average of the ethnic group SAT–M means after adjusting for the other predictors of SAT–M. The estimate of 2.7 for ETH3 indicates that the Hispanic group is 2.7 points below the average of the ethnic group SAT–M means after adjusting for the other predictors of SAT–M. We can transform these estimates into values that reflect the difference between the white group and each of the other ethnic groups, again after adjustment for the other predictors in the model. This was done by summing all the contrast variables—ETH1, ETH2, and ETH3—and then adding the sum to the path coefficient estimates for ETH1, ETH2, and ETH3. The resulting values show that the difference between the white and African American adjusted SAT–M means is 32.9 points, the difference between white and Asian adjusted means is 4 points, and the difference between white and Hispanic adjusted means is 15.9 points. In all of these differences, the white group has the highest scores. Only the difference between whites and African Americans is meaningful in size, with this difference being approximately a third of a standard deviation on the SAT–M. Finally, the ethnic-by-gender interactions (SETH1, SETH2, SETH3) were either not statistically significant, or were not meaningful in size.

For SAT–V, the pattern of results is similar except that the gender difference is negligible (13.2 points on the SAT–V scale). The difference between whites and African Americans is 30.8 after adjustment for other predictors of SAT–V. The adjusted difference between whites and Asian Americans on SAT–V is 28.9 points, a larger difference between these groups than was found for SAT–M. The adjusted difference for whites and Hispanics is 13.5. The ethnic/gender interactions on SAT–V are again negligible. In summary, the largest effects found were for gender on SAT–M, followed by the difference between whites and African Americans on both the SAT–M and SAT–V, and the difference between whites and Asian Americans on the SAT–V.

Moving to the remaining portion of the within-school structural model, Table 10 gives the standardized path coefficient estimates for the remaining direct paths.

All estimates are statistically significant. Each estimate has the following interpretation: Adjusting for all other predictors, the estimate gives the expected difference in the criterion variable, in standard deviation units, for a difference of one standard deviation on the predictor. For example, for the path from family background to SAT–M, the estimate is .047. A difference of one standard deviation on family background is equivalent to a .047 standard deviations difference on the SAT–M, after adjusting for other predictors of SAT–M. The meaning of a standard deviation on the SAT–M is clear, but a difference of one standard deviation on family background is less clear because it is a latent variable and is scaled as such. To understand the meaning of a standard deviation on family background, we must refer back to the measurement portion of the within-school model. Table 11 gives the factor-loading estimates, standardized to correspond to latent variable variances of 1.0.

To illustrate, the loading estimate for HNRCNT under extracurricular activities is 1.285. This means that an increase of one standard deviation on the extracurricular activities latent variable is equivalent to 1.285 more honors courses. Generally, a loading estimate in Table 11 indicates the expected difference on the observed variable's scale for one standard deviation's difference in the factor or latent variable. A standard deviation increase in family background translates into (1) a 1.6-point increase on the FATHED scale, (2) a 1.506-point increase on the FAMINC scale, and (3) a 1.338-increase on the MOTHED scale. It is clear, however, that the direct impact of family background on both SAT–M and SAT–V is negligible, given the path coefficient estimates in Table 10. On the other hand, the direct paths from high school achievement to SAT–M, and the paths from extracurricular activities to SAT–M and SAT–V, appear to be substantial. In each case, the expected change in SAT performance ranges from just less than one-third of a standard deviation to one-half of a standard deviation on the SAT scale.

Table 10

Standardized Path Coefficient Estimates for Latent Predictors of SAT for the Within-School Model

	High School Achievement	*Extracurricular Activities*	*SATM*	*SATV*
Family Background	.257	.333	.047	.079
High School Achievement			.309	.155
Extracurricular Activities			.445	.536

Table 11

Standardized Factor-Loading Estimates for the Within-School Model

	High School Achievement	*Extracurricular Activities*	*Family Background*
SCIGR	.496		
HSAVG	-1.703		
CRANK	-.949		
ARTGR	.197		
ENGR	.458		
LANGR	.550		
MATHGR	.539		
SOCGR	.441		
ACTCNT		.965	
HNRCNT		1.285	
APCNT		.927	
ARTCNT		.319	
COMPCNT		.546	
ENGCNT		.728	
FATHED			1.600
FAMINC			1.506
MOTHED			1.338

Note: Estimates are standardized to correspond to factor variances of 1.0.

To gain perspective on this, a standard deviation change on the high school achievement latent variable, interpreted using the loading estimates in Table 11, corresponds to (1) a 1.703 change on the HSAVG scale (see the Appendix) and (2) a .949 change in the CRANK scale, with expected changes in the remaining scales calculated similarly. For the extracurricular activities latent variable, a standard deviation change on the latent scale corresponds to (1) 1.285 additional honors courses (HNRCNT) and (2) .927 additional AP® courses that the student intends to take (APCNT), with expected changes in the other scales calculated analogously. Table 10 also reveals substantial direct paths between family background and both the high school achievement and extracurricular activities latent variables. We can infer from these results that there is an indirect link between family background and SAT performance that is mediated by these two latent variables.

Table 12 presents the path coefficient estimates for the paths from ethnic/gender contrasts to high school

Table 12

Standardized Estimates for Ethnic-by-Gender Paths to Latent Variables for the Within-School Model

	High School Achievement	*Extracurricular Activities*
SEX	-.146	-.119
ETH1	.438	.423
ETH2	-.375	-.434
ETH3	.004*	-.010*
SETH1	.082	.061
SETH2	-.033	-.024
SETH3	-.053	-.019

*Not statistically significant at alpha = .05.

achievement and extracurricular activities. These estimates have been standardized to correspond to latent variable variances of 1.0. Hence they express group differences on the latent variable in standard deviation units on the latent scale.

The interaction paths (SETH1, SETH2, SETH3) are negligible in every case, as are the paths from ETH3, which involved the Hispanic group. The SEX paths show that females have higher scores than males on both school achievement and extracurricular activities latent variables, adjusting for other predictors, though the differences are relatively small. The largest differences are found for the African Americans (ETH1) and the Asian Americans (ETH2). Adjusting for other predictors, African Americans score a little more than four-tenths of a standard deviation below the mean on both the high school achievement and extracurricular activities latent measures. The "mean" here is the average of the four ethnic group means on the respective outcome. In contrast, the Asian American group scores close to four-tenths of a standard deviation above the mean on both these latent variables, adjusting for other predictors. Finally, the within-school model provides estimates of the squared-multiple correlations for each endogenous variable, given the predictors of that variable as modeled. These estimates are: SAT–M = .570, SAT–V = .492, high school achievement = .112, and extracurricular activities = .145.

The Between-School Model. Turning now to the between-school modeling results, Table 13 presents the path coefficient estimates for the structural model. The estimates are standardized to correspond to latent-variable variances of 1.0.

To aid in the interpretation of the latent-variable scales, Table 14 presents the factor-loading estimates from the between-school model for the three latent variables. These estimates are also standardized to correspond to latent-variable variances of 1.0. In interpreting any of

Table 13

Standardized Path Coefficient Estimates for the Between-School Model

	High School Achievement	*Extracurricular Activities*	*SAT–M*	*SAT–V*
MEMBER	-.046	-.128	1.080*	.860*
FLE	.160	.216	-.380*	.003*
LOCALE	.033	-.042	-.960*	-.340*
PNWHIT	-.082	-.086	-7.590	-8.170
Family Background	.463	.881	60.10	54.60
High School Achievement			8.90	3.60
Extracurricular Activities			20.80	34.00

*Not statistically significant at alpha = .05.

the parameter estimates in Tables 13 and 14, it must be remembered that the unit of analysis is the school, and any measures refer to school-level averages.

From Table 14, the highest-loading variables on the high school achievement factor are SCIGR, SOCGR, and HSAVG. A difference of one standard deviation on high school achievement is equivalent to (1) a 1-point difference on the grade scale for SCIGR; (2) a .982-point difference on the grade scale for SOCGR; and (3) a .592-point difference on the HSAVG scale. The

Table 14

Standardized Factor-Loading Estimates for the Between-School Model

	High School Achievement	*Extracurricular Activities*	*Family Background*
SCIGR	1.00		
HSAVG	-.592		
CRANK	-.190		
ARTGR	.095		
ENGR	.187		
LANGR	.168		
MATHGR	.195		
SOCGR	.982		
ACTCNT		.514	
HNRCNT		.099	
APCNT		.162	
ARTCNT		.201	
COMPCNT		.209	
ENGCNT		.281	
FATHED			.938
FAMINC			1.344
MOTHED			.683

Note: Estimates are standardized to correspond to factor variances of 1.0.

highest-loading variables on the extracurricular activities factor are ACTCNT and ENGCNT. A difference of one standard deviation on this latent variable is equivalent to (1) a .514-point difference on additional activities in ACTCNT; and (2) a .281-point difference on additional activities in ENGCNT. A difference of one standard deviation on the family background is equivalent to (1) a .938-point difference on the FATHED; (2) a .683-point difference on the MOTHED scale; and (3) a 1.344-point difference on the FAMINC scale. As an additional point of interpretation, the path coefficients for MEMBER and FLE school-effect variables are scaled so that a unit on each of these variables refers to 100 students. This rescaling was necessary for numerical stability in the estimation. For ease of interpretation, the path coefficients for PNWHIT are scaled so that a unit on PNWHIT is a 10 percent difference in the percentage of minority students.

With this information as background, we can more easily interpret the path coefficient results found in Table 13. First, with the exception of percentage of minority students enrolled (PNWHIT), none of the between-school-level variables in our models are related directly to SAT performance. The results for PNWHIT, however, show that after adjusting for other predictors of SAT performance, a 10 percent increase in the percentage of minority students is associated with about an 8-point drop in average SAT performance, considering both SAT–M and SAT–V. This effect is not large in comparison to the coefficient estimates for the latent predictors of family background and extracurricular activities. A standard deviation increase in family background, for example, is associated with a 60-point increase in SAT–M, and a 55-point increase in SAT–V, adjusting for the other predictors in the model. (Note here that "family background" denotes the average family background for students in the school who took the SAT.) Changes in the family background latent variable in the between-school model, therefore, denote shifts in the average family background status among the students in these schools. The same comment applies to both the high school achievement and extracurricular activities factors. High school achievement is not associated with a large change in average SAT performance: only about 9 points on SAT–M and about 4 points on the SAT–V per standard deviation change in high school achievement. However, the extracurricular activities factor is associated with substantial changes in average SAT performance: about 21 points on SAT–M and 34 points on SAT–V, per standard deviation change in the extracurricular activities factor.

Turning to the direct paths to the two latent variables of high school achievement and extracurricular activities, it is clear that family background is associated with substantial change in both. Adjusting for other predictors, an increase of one standard deviation on family background, we learned, is associated with almost a one-half standard deviation change in average high school achievement, and nearly nine-tenths of a standard deviation change in average extracurricular activities. The latter result implies that family background has an indirect effect on SAT performance that is mediated by the latent variable of extracurricular activities. Among the school-level variables, LOCALE had only a negligible direct relation to high school achievement and extracurricular activities. The impact of MEMBER, i.e., the size of the high school, is also relatively small, as a 100-student increase in enrollment is associated with less than one-tenth of a standard deviation decrease in SAT–M, and about 13 percent of one standard deviation decrease in SAT–V. The effects found for FLE, the number of students eligible for free and reduced lunch, are also relatively small, but the effects' direction is surprising: An increase in the number of free-lunch students is associated with an increase in SAT performance. We return to this finding below.

Finally, PNWHIT shows a substantial effect on both high school achievement and extracurricular activities. An increase of 10 percent in PNWHIT is associated with a decrease of between 80 percent and 90 percent of a standard deviation in both high school achievement and extracurricular activities. The between-school model provides estimates of the squared-multiple correlations for each endogenous variable (high school achievement, extracurricular activities, SAT–M, SAT–V). These estimates are: high school achievement = .165, extracurricular activities = .655, SAT–M = .797, and SAT–V = .858. The R^2 values for SAT–M and SAT–V are considerably higher here than in the within-school model, reflecting the more predictable behavior of school averages.

Discussion

Throughout this study we used multilevel structural equation models to explore the effects of both individual differences and school effects on SAT scores. Though complex, these models nevertheless are revealing and point to a number of important directions for future research on large-scale test performance, efforts that go beyond examining individual differences, or group differences, for that matter. However, before discussing the implications of these analyses, it is important to call attention to the methodological limitations of this study.

First, and most obvious, both the high school and SAT data were not gathered, necessarily, to answer the kinds of questions we raise in this study. We would be remiss, then, if we did not acknowledge the possibility that other, theoretically more important, indices are missing from our models. This is particularly obvious when it comes to the school-level measures. Unfortunately, other, perhaps more powerful, explanatory measures were unavailable to us. Similarly, missing data at the student level were handled by excluding those cases with missing data—not the most desirable method. Future research ought to test the viability of using multilevel models that allow for more robust imputation techniques to handle the possible biases related to nonrandom patterns of missing data.

Second, we also recognize the problems associated with students' nonrandom assignment to schools in the public sector. To address this concern, we attempted to closely match students based on a host of individual differences in family background, school achievement, and academic experiences. In the absence of randomized field studies, it is difficult to provide stronger evidence of school effects. The use of multilevel, latent-variable models, however, does point to empirically derived potential evidence in these areas.

Finally, the student-level model tested here assumes a linear relationship among and between the family background measures, school achievement, school activities, and performance on the SAT. This is largely speculative, and the relationships may not be as direct and uniform as our models imply. Further research with nonlinear models ought to examine these assumptions.

Despite these obvious limitations, the analyses and findings of this study highlight the considerable advantages of using multilevel-latent-variable models to understand student achievement—particularly when compared with more conventional, ordinary, least-squares regression methods. As we noted earlier, latent-variable models allow for a more complete, more complex picture of the relationships among and between hypothesized predictors and outcomes. Single observed variables, like those used in regression models, are largely inadequate and often more unreliable than latent measures. The use of latent variables, therefore, provides a more nuanced understanding of the complexities of performance on standardized tests.

Of singular importance, we suspect, is the advantage of avoiding the ecological fallacy (Robinson, 1950; Snijders and Bosker, 1999), the correlation between second-level (or school-level) variables used, albeit often inadvertently, to make assertions about individual-difference-level variables. The percentage of minority students in a school, for example, could be related to the average test scores in that school. However, this correlation provides little understanding about the individual differences, say, between ethnicity and academic achievement. Multilevel models, therefore, allow us to disentangle the correlations between macrolevel and microlevel variables.

Summary

At the student level, the nested series of models we tested suggest that the five-factor measurement model with invariant factor loadings fits the data reasonably well. Thus, we are confident the measurement model provides a relatively robust method of condensing a large and unwieldy number of individual difference measures to a smaller, more elegant, and theoretically useful set of latent variables. In the process we gain a measurable and appreciable amount of explanatory power. The fit of the latent-variable model is reasonably good across the subgroups—though more research is needed on the models as they apply to African American students, both males and females.

These student-level models, moreover, shed light on the relative importance of extracurricular activities on high-stakes tests. Like other investigators (Camp, 1990; Gerber, 1996; Holloway, 2000; Marsh and Kleitman, 2002), our study provides compelling evidence from the SAT that participation in extracurricular activities provides all students—including students from disadvantaged backgrounds, minorities, and those with otherwise less-than-distinguished academic achievements in high school—a measurable and meaningful gain in their college admissions test scores. The important reasoning abilities measured by tests like the SAT, evidently, are developed both in and out of the classroom. To paraphrase Marsh and Kleitman (2002), participation in extracurricular activities in high school appears to be one of the few interventions that may benefit disadvantaged students—those less well served by traditional educational programs—as much as or more than their more advantaged peers.

On the other hand, the oft-cited relationship between family wealth and socioeconomic background and SAT scores, at the individual student level, appears to be moderated by both student achievement levels and exposure to extracurricular activities. This is not to say that family background—particularly parental education levels—does not matter. But these models suggest that the relationship is complex and moderated by school resources, as well as family assets.

At the same time, the structural models at the student level, though useful and informative, were not entirely invariant across racial/ethnic groups. The exception was the relatively poor fit of the model to the data from the African American students. Obviously, our models for African American students are inadequate and require more work. At the very least, they need to be expanded to include variables and indicators of the quality of the high schools with large proportions of African American students, as well as to be informed by affective and other variables that capture levels of academic engagement (Gordon, 1999). Given the historical patterns of racial segregation in housing and educational opportunities in the United States, it would be surprising, indeed, if one generic model would fit this particular group of African American students.

Conclusion

Our analyses shed considerable light on the influence of school effects. Put succinctly, these models demonstrate the importance of the high school and the schools' contexts (Lee, 2000). Evidence from no less an important measure as the SAT provides strong support for the claim that schools, and the differences between them, matter. Paraphrasing Lee (2000), our modeling efforts make clear that the structure and organization of high schools—even when looking only at publicly funded schools—influence student achievement. School size, the proportion of children in poverty, and the ethnic/racial composition of the high schools were all important and meaningful predictors of student achievement, beyond the individual differences that children bring with them to the schools. Clearly, the work we report here echoes research reported earlier by our colleague Valerie Lee.

> Children's learning is strongly influenced by the contexts in which it occurs. Those contexts may be defined by the children's families, the classmates with whom they experience schooling, the peers with whom they choose to interact, and the teachers who instruct them. Students are profoundly influenced by the schools they attend (Lee, 2000, p. 140).

The application of multilevel structural modeling techniques to data from the College Board's SAT was revealing. Advocates of school reform have, at times, attacked high-stakes tests such as the SAT as biased, unfair, or discriminatory. The analyses reported here provide little support and comfort to those critics. On the contrary, most of the between-group differences in SAT scores shrink to negligible levels (many well within the standard errors of measurement of both the SAT–V and SAT–M tests), once other influences at both the student- and school-levels are included in the analyses. Again, the central point emerging from our analyses is that context matters—more simply, schools matter when it comes to promoting differences in student achievement.

We hope that at the very least our work can serve to animate research in education and psychology to move beyond individual differences, and beyond the traditional two disciplines of experimental and correlational research on student learning (Cronbach, 1957). Our intent is for these modeling approaches to become more widely used and, in the process, to further research that goes beyond studies of variance within and between students ("organisms," in Cronbach's language). Our vision is that these modeling approaches will provide a means for unifying research design so as to better address the interactions of students as learners, educational programs as treatments, and schools as the contexts in which they occur.

References

Anderson, B.T., & Bruschi, B.A. (February 1993). *Gender differences in mathematics achievement between and within minority-majority subgroups.* Paper presented at the annual meeting of the Eastern Educational Research Association, Clearwater, FL.

Bentler, P.M. (1990). "Comparative fit indexes in structural models." *Psychological Bulletin*, *107*, 238–246.

Bentler, P.M., & Bonett, D.G. (1980). "Significance tests and goodness of fit in the analysis of covariance structures." *Psychological Bulletin*, *88*, 588–606.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Camara. W.J., & Schmidt, A.E. (1999). *Group differences in standardized testing and social stratification*. College Board Report No. 99–5. New York: College Entrance Examination Board.

Camp., W. (1990). "Participation in student activities and achievement: A covariance structural analysis." *Journal of Educational Research*, *83*, 272–278.

Cohen, D.K., Raudenbush, S.W., & Ball, D.L. (2002). "Resources, instruction, and research." In F. Mostellar & R. Boruch (eds.), *Evidence matters: Randomized trials in educational research* (pp. 80–119). Washington, DC: Brookings Institution Press.

Coleman, J. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

College Board (2003). *College-Bound Seniors*. New York: College Entrance Examination Board.

Cronbach, L.J. (1957). "The two disciplines of scientific psychology." *American Psychologist*, *30*, 116–127.

Dabady, M. (2003). "Measuring racial disparities and discrimination in elementary and secondary education: An introduction." *Teachers College Record*, *105(6)*, 1048–1051.

Dorans, N.J. (2002). *The recentering of SAT scales and its effects on score distributions and score interpretations.* New York: College Entrance Examination Board, No. 2002–11.

Everson, H.T., & Michna, G. (January 2004). *Is the SAT a wealth test? The influence of family income, ethnicity and high school achievement on SAT scores.* Paper presented at a colloquium sponsored by the Institute for Urban and Minority Education, Teachers College, Columbia University, New York, NY.

Everson, H.T., Millsap, R.E., & Diones, R. (April 1995). *Exploring group differences in SAT performance using structural equation modeling with latent means*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Gerber, S. (1996). "Extracurricular activities and academic achievement." *Journal of Research and Development in Education*, *30(1)*, 42–50.

Goldstein, H. (1987). *Multilevel models in educational and social research*. New York: Oxford University Press.

Good, T.L., & Brophy, J.E. (1986). "School effects." In M.C. Wittrock (ed.), *Handbook of research on teaching* (pp. 570–602). New York: Macmillan.

Gordon, E.W. (1999). *Education and justice: A view from the back of the bus*. New York: Teachers College Press.

Hanushek, E.A. (1997). "Assessing the effects of school resources on student performance. An update." *Educational Evaluation and Policy Analysis*, *19(2)*, 141–164.

Holloway, J.H. (2000). "Extracurricular activities: The path to academic success." *Educational Leadership*, *57* (*4*), 87–88. Arlington, VA: Association for Supervision and Curriculum Development.

Hoyle, R.H. (ed.) (1995). *Structural equation modeling: Concepts, Issues, and Applications*. Thousand Oaks, CA: Sage.

Huang, G., Reiser, M., Parker, A., Muniec, J. & Salvucci, S. (2003). *Institute of education sciences findings from interviews with education policymakers*. Arlington, VA: Synectics.

Jencks, C., & Phillips, M. (1998). *The black–white test score gap*. Washington, DC: The Brookings Institute Press.

Jones, L.V. (1987). "The influence on mathematics test scores, by ethnicity and sex, of prior achievement and high school mathematics courses." *Journal for Research in Mathematics Education*, *18(3)*, 180–186.

Kozol, J. (1991). *Savage inequalities: children in America's schools*. New York: Crown Publishers.

Lee, V.E. (2000). "Using hierarchical linear modeling to study social contexts: The case of school effects." *Educational Psychologist*, *35(2)*, 125–141.

Lee, V.E., Bryk, A.S., & Smith, J.B. (1993). "The organization of effective high schools." In L. Darling-Hammond (ed.), *Review of Research in Education*, *19*, 171–267. Washington, DC: American Educational Research Association.

Loehlin, J.C. (1992). *Latent variable models*. Hillsdale, NJ: Lawrence Erlbaum.

Longford, N.T. (1993). *Random coefficient model*. New York: Oxford University Press.

Marsh, H. (1992). "Extracurricular activities: Beneficial extension of the traditional curriculum or subversion of academic goals?" *Journal of Educational Psychology*, *84(4)*, 553–562.

Marsh, H.W., & Kleitman, S. (2002). "Extracurricular school activities: The good, the bad, and the nonlinear." *Harvard Educational Review*, *72*, 4, 464–514.

McDonald, R.P. (1993). "A general model for two-level data with responses missing at random." *Pyschometrika*, *4*, 575–585.

McDonald, R.P., & Marsh, H.W. (1990). "Choosing a multivariate model: Noncentrality and goodness of fit." *Psychological Bulletin*, *107*, 247–255.

Mickelson, R.A. (2003). "When are racial disparities in education the result of racial discrimination? A social science perspective." *Teachers College Record*, *105*, 6, 1052–1086.

Millsap, R.E., (2002). "Structural equation modeling: A user's guide." In F. Drasgow & N. Schmitt (eds.), *Measuring and analyzing behavior in organizations*. (pp. 257–301). San Francisco: Jossey-Bass.

Millsap, R. E., & Everson, H. (1991). "Confirmatory measurement model comparisons using latent means." *Multivariate Behavioral Research*, *26(3)*, 479–497.

Muthen, B. (1984). "A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators." *Psychometrika*, *49*, 115–132.

Muthen, L.K., & Muthen, B. (1998). *Mplus user's guide*. Los Angeles: Muthen & Muthen.

National Center for Education Statistics (1994). *The Common Core Data Public/Elementary School Universe Survey: School Year 1993-94* (NCES Publication No. 1994333). Washington, DC: Author.

National Center for Education Statistics (2002). *The Conditions of Education*. Washington, DC: U.S. Department of Education.

Raudenbush, S.W., & Sampson, R. (1999). "Assessing direct and indirect effects in multilevel designs with latent variables." *Sociological Methods & Research*, *28(2)*, 123–153.

Robinson, W.S. (1950). "Ecological correlations and the behavior of individuals." *American Sociological Review*, *15*, 351–357.

Snijders, T.A.B., & Bosker, R.J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage Publications, Ltd.

Steiger, J.H., & Lind, J.M. (1980). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

U.S. Department of Education, NCES (2002). *The Conditions of Education*. (NCES 2002–025). Washington, DC: U.S. Government Printing Office.

Appendix

Table A1. Means of Measured Variables by Race/Ethnicity
Males

Variables	*Whites*	*Asian Americans*	*African Americans*	*Hispanics*
HSAVG	4.33	3.98	5.56	4.66
CRANK	2.65	2.51	3.15	2.87
ARTGR	3.66	3.72	3.42	3.59
SOCGR	3.34	3.42	2.97	3.22
ENGR	3.16	3.27	2.81	3.06
LANGR	3.05	3.25	2.67	3.17
MATHGR	3.10	3.27	2.65	2.93
SCIGR	3.21	3.31	2.78	3.04
ACTCNT	2.96	2.80	2.26	2.27
APCNT	1.25	1.92	0.94	1.19
HNRCNT	1.46	1.95	0.85	1.36
ENGCNT	4.11	3.98	3.16	3.66
COMPCNT	2.81	2.98	2.20	2.50
ARTCNT	1.66	1.65	1.38	1.52
FATHED	6.15	6.43	5.01	4.77
MOTHED	5.69	5.86	5.28	4.49
FAMINC	8.82	8.36	6.31	6.67
SAMPLE SIZE	**170,270**	**12,333**	**18,411**	**13,026**

Table A2. Means of Measured Variables by Race/Ethnicity
Females

Variables	*Whites*	*Asian Americans*	*African Americans*	*Hispanics*
HSAVG	3.83	3.55	4.83	4.30
CRANK	2.51	2.41	2.97	2.85
ARTGR	3.82	3.83	3.61	3.70
SOCGR	3.40	3.48	3.11	3.26
ENGR	3.42	3.49	3.11	3.27
LANGR	3.35	3.51	3.05	3.38
MATHGR	3.12	3.23	2.71	2.88
SCIGR	3.25	3.34	2.91	3.06
ACTCNT	3.33	3.26	2.56	2.51
APCNT	1.21	1.84	1.01	1.16
HNRCNT	1.63	2.18	1.21	1.54
ENGCNT	4.36	4.23	3.59	3.94
COMPCNT	2.59	2.75	2.52	2.52
ARTCNT	2.13	2.11	1.75	1.88
FATHED	5.90	6.29	4.70	4.57
MOTHED	5.52	5.71	5.03	4.37
FAMINC	8.44	8.06	5.70	6.33
SAMPLE SIZE	**212,412**	**13,732**	**27,644**	**16,666**

CollegeBoard
connect to college success™

www.collegeboard.com

040481303